



Title	A Close-up Look at the Chemical Space of Commercially Available Building Blocks for Medicinal Chemistry
Author(s)	Zabolotna, Yuliana; Volochnyuk, Dmitriy M.; V. Ryabukhin, Sergey; Horvath, Dragos; Gavrilenko, Konstantin S.; Marcou, Gilles; Moroz, Yuri S.; Oksiuta, Oleksandr; Varnek, Alexandre
Citation	Journal of chemical information and modeling, 62(9), 2171-2185 <a href="https://doi.org/10.1021/acs.jcim.1c00811">https://doi.org/10.1021/acs.jcim.1c00811</a>
Issue Date	2021-12-20
Doc URL	<a href="http://hdl.handle.net/2115/87610">http://hdl.handle.net/2115/87610</a>
Rights	This document is the Accepted Manuscript version of a Published Work that appeared in final form in Journal of chemical information and modeling, copyright c American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work see <a href="http://pubs.acs.org/articlesonrequest/AOR-819JU9NYIEPICSWP4865">http://pubs.acs.org/articlesonrequest/AOR-819JU9NYIEPICSWP4865</a> .
Type	article (author version)
File Information	ZABOLOTNA_BBs_FINALsubmission.pdf



[Instructions for use](#)

# A CLOSE-UP LOOK AT THE CHEMICAL SPACE OF COMMERCIALLY AVAILABLE BUILDING BLOCKS FOR MEDICINAL CHEMISTRY

*Yuliana Zabolotna<sup>1</sup>, Dmitriy M.Volochnyuk<sup>3,6</sup>, Sergey V.Ryabukhin<sup>4,6</sup>, Dragos Horvath<sup>1</sup>, Konstantin S.Gavrilenko<sup>5,6</sup>, Gilles Marcou<sup>1</sup>, Yurii S.Moroz<sup>5,7</sup>, Oleksandr Oksiuta<sup>3,7</sup>, Alexandre Varnek<sup>1,2\*</sup>*

1. University of Strasbourg, Laboratoire de Chemoinformatique , 4, rue B. Pascal, Strasbourg 67081 (France) \*e-mail: [varnek@unistra.fr](mailto:varnek@unistra.fr)
2. Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan
3. Institute of Organic Chemistry, National Academy of Sciences of Ukraine, Murmanska Street 5, Kyiv 02660, Ukraine
4. The Institute of High Technologies, Kyiv National Taras Shevchenko University, 64 Volodymyrska Street, Kyiv 01601, Ukraine
5. Research-And-Education ChemBioCenter, National Taras Shevchenko University of Kyiv, Chervonotkatska str., 61, 03022 Kiev, Ukraine
6. Enamine Ltd. 78 Chervonotkatska str., 02660 Kiev, Ukraine
7. Chemspace, Chervonotkatska Street 78, 02094 Kyiv, Ukraine.

**ABSTRACT.** The ability to efficiently synthesize desired compounds can be a limiting factor for chemical space exploration in drug discovery. This ability is conditioned not only by the existence of well-studied synthetic protocols but also by the availability of corresponding reagents, so-called building blocks (BB). In this work, we present a detailed analysis of the chemical space of 400,000 purchasable BB. The chemical space was defined by corresponding synthons – fragments

contributed to the final molecules upon reaction. They allow an analysis of BB physicochemical properties and diversity, unbiased by the leaving and protective groups in actual reagents. The main classes of BB were analyzed in terms of their availability, rule-of-two-defined quality and diversity. Available BBs were eventually compared to a reference set of biologically relevant synthons derived from ChEMBL fragmentation, in order to illustrate how well they cover the actual medicinal chemistry needs. This was performed on a newly constructed universal generative topographic map of synthon chemical space, that enables visualization of both libraries and analysis of their overlapped and library-specific regions

## INTRODUCTION

The success of drug discovery strongly depends on the quality of the screening compounds. Starting molecules may be derived from natural sources or synthesized by organic chemists. Even though natural products have been evolutionarily selected to bind to biological macromolecules efficiently, they may not be easy to extract and purify on a large industrial scale. The pursuit of structural diversity with easily obtainable compounds led to the mutually dependent symbiotic relationships between drug discovery and organic synthesis.<sup>1</sup>

Over the past decades, the chemical market has evolved to meet medicinal chemistry demands, with new compounds having medicinal chemistry relevant physicochemical properties – low molecular weight and lipophilicity, high fraction of sp<sup>3</sup> carbon atoms (Fsp<sup>3</sup>), etc.<sup>2</sup> At the same time, it is well known that chemotype distribution in the commercially available libraries of screening compounds is highly unbalanced towards synthetically accessible benzenesulfonamides, anilides, and other amides, etc.<sup>3</sup> Beyond the immediately available "on-shelf" collections, "tangible libraries" of easily accessible (but not yet produced) molecules were proposed.<sup>4</sup> They have emerged as the result of the stock enhancement campaigns, directed towards the overall improvement of collections' quality and novelty. However, they still tend to sample already overpopulated areas of the chemical space.<sup>3</sup> That means that the commercial library enhancement strategies do not provide a uniform chemical space sampling, and thus there is an urgent need for their improvement.

One of the most efficient ways to do that consists of early quality control via monitoring properties and novelty of used building blocks (BB) or reagents - molecules that possess at least

one reactive functional group and thus can participate in the synthesis of the final screening molecules. Usage of the medicinally relevant BB can significantly improve the quality of the designed compounds by preliminary focusing on substructures and properties that will ensure desirable activity and the ADMETox profile of the potential drug candidates. Moreover, introducing the new BB will allow exploration of underrepresented regions of the chemical space, potentially accessing various properties and bioactivities.

Even though medicinal chemists widely recognize this fact, the number of scientific reports targeting quality analysis of the existing purchasable building blocks (PBB) and potential strategies for the corresponding libraries enhancement is significantly lower than those concerning commercially available screening compounds. Within the last two decades, the latter has been evaluated in numerous medicinal chemistry publications.<sup>2, 3, 5-11</sup> At the same time, there are only a few works dedicated to BB used in medicinal chemistry.

Based on the AstraZeneca (AZ) five-year 'long strategic reagent initiative', Goldberg et al.<sup>12</sup> outlined general design principles for novel BB in order to maximize their impact on drug discovery projects. Besides, they listed the most popular types of BB, chosen by medicinal chemists from AstraZeneca for different drug design campaigns. In another study, Hartenfeller et al.<sup>13</sup> investigated the biological relevance of the chemical space spanned by 58 of the most popular organic chemistry reactions, based on a subset of the readily available BB ( $\approx 26,000$ ). They have concluded that established synthetic resources are well suited to cover selected biologically relevant compounds. However, the chosen reference subset was limited to only  $\approx 62,000$  compounds from GVK-BIO<sup>14</sup>, Drug Bank<sup>15</sup>, and TIMBAL<sup>16</sup>, which might fall short as a comprehensive representation of all known biologically active compounds.

Moreover, the analysis of all PBB was beyond the scope of both mentioned papers. To our best knowledge, the only report of such analysis is a price-focused study of almost one million PBB from 121 vendors, published by Kalliokoski.<sup>17</sup> In this work, he analyzed the availability of the 13 types of BB, reporting the number of reagents available for purchase under a specific range of price up to \$150/g. However, even though all these reports provide an important insight into the PBB libraries and some of the medicinal chemistry relevant properties, those articles, each being published at least five years ago, can hardly characterize the current state of the quickly growing chemical space the PBB.

Therefore, in this work, we present the analysis of an up-to-date PBB set, addressing the availability of the most popular classes of BB, their diversity, and their ability to face current medicinal chemistry needs in the synthesis of biologically relevant compounds. As a source of PBB in-stock database of the biggest BB aggregator, eMolecules Inc.<sup>18</sup> has been used. For their analysis, we have employed the previously reported freely available python library – Synthons Interpreter (SynthI) – knowledge-based reaction toolkit for the library analysis and design.<sup>19</sup> It allows examining BB not as individual chemical entities but as a set of synthons – fragments obtained after leaving groups removal/transformation with a system of labels that encodes position and type of reactive center (RC). They define the substructure that will contribute to the final molecule upon different reactions (except heterocyclization, omitted in this analysis). The same tool has been used for fragmenting compounds from ChEMBL<sup>20</sup> in order to detect synthons and, if available, corresponding BB required for the synthesis of the biologically relevant molecules from this database. The diversity of synthons has been analyzed using marked-atom ISIDA fragment descriptors<sup>21</sup> that consider the marked connection points in synthons (former locations of the leaving groups). Thanks to that, it becomes possible to distinguish between BB that structurally differ only in terms of leaving groups and RC placement. These descriptors were also used to define the chemical space of BB, which was visualized via Generative Topographic Mapping (GTM).<sup>22</sup> This non-linear visualization method has proven multiple times to be effective in the analysis of large chemical databases.<sup>3, 23-27</sup> However, it is the first time it was used to map the space of synthons.

## DATA

The 489,781 building blocks, provided by eMolecules, Inc.,<sup>18</sup> have been used as a source of readily available PBB. Unique chemical structures within Tier 1 or 2 (corresponding to shipments within 5 and 10 days, respectively) were selected to represent in-stock compounds.

ChEMBL (version 26) served as a reference dataset for biologically relevant molecules. 1,950,765 compounds have been standardized according to the procedure implemented on the virtual screening server of the Laboratory of Chemoinformatics (infochimie.u-strasbg.fr/webserv/VSEngine.html), using the ChemAxon Standardizer<sup>28</sup>. That included:

- dearomatization and final aromatization (heterocycles like pyridone were not aromatized);

- conversion to canonical SMILES;
- salts and mixture removal; neutralization of all species, except nitrogen (IV);
- major tautomer generation
- stereochemical information removal.

Stereochemical information has been ignored because used ISIDA descriptors<sup>21</sup> would not capture it, anyway. 1,721,155 unique ChEMBL compounds, remaining after standardization, were then fragmented in order to obtain biologically relevant synthons. The resulting synthons, as well as synthons generated from eMolecules library, were standardized according to the same procedure.

## METHODS

### Synthons Interpreter (SynthI)

Considering that a single BB can contribute different structural motifs to the molecule, depending on the synthesis conditions and reaction partners, it is not helpful to analyze the primary chemical structures of the BB in the context of their usage in medicinal chemistry. Different protective and leaving groups can constitute a large (sometimes the largest) part of the reagent. Synthons, by contrast, represent the substructure of a BB that will be inherited by the product, annotated by marks on the atoms that will connect to partner synthons. In our previous work, we have developed a python library - Synthons Interpreter (SynthI) - for synthon generation from either BB or drug-like products by RECAP-based fragmentation.<sup>19</sup> The code is freely available at <https://github.com/Laboratoire-de-Chemoinformatique/SynthI>. It consists of four modules, three of which were used in this work:

1. **SynthI-Classifier** consists of the library of SMARTS identifying structural motifs required and respectively forbidden in BB suitable as a particular class of reagents required by the considered set of chemical reactions. For now, this set only includes coupling reactions (no heterocyclization reactions). These involve 22 generic monofunctional reagent classes, like acyl halides, boronics, ketones, primary amines, etc. These can be further subdivided into about 100 finer subclasses of significantly diverging reactivities. For example, class "Alcohols" includes three subclasses of reactivity – "Heterols", "Aliphatic alcohols", and "Phenols". In addition, there are 28 bifunctional and 19 trifunctional classes.

2. **SynthI-BB** allows generation of the most probable synthons exhaustively from a given BB – a process herein referred to as "synthonization". The position of the functional groups, as well as the formal type of the resulting fragment (electrophilic, nucleophilic, radical, etc.), is encoded as synthons with class-specific numeric marks on the "connecting" atoms. There are nine types of RC that can appear in synthons:

- *electrophilic* (produced by acyl and aryl halides, acids, aldehydes, ketones, etc.);
- *nucleophilic* (alcohols, thiols, amines, metal organics, hydrazines, hydrazides etc.);
- *bivalent electrophilic* (aldehydes and ketones);
- *bivalent nucleophilic* (primary amines, hydroxylamines, reagents for olefination, etc.);
- *bivalent neutral* (terminal alkenes for metathesis);
- *electrophilic radical* (Minisci CH-partners, Michael acceptors);
- *nucleophilic radical* (BF<sub>3</sub> and MIDA boronates, NOPhtal alkyl esters, sulphinates, etc.);
- *boronics-derived nucleophilic* (boronic reagents);
- *electrophilic nitrogen* (benzoyl O-acylated hydroxylamines).

Examples can be found in **Table S1** of Supporting Information (SI).

The resulting synthons, represented by ISIDA descriptors, were used to define the chemical space of commercially available BB. The type of ISIDA fragments was selected during GTM optimization (see further).

3. **SynthI-Fragmentation** was used to evaluate current PBB space's ability to face medicinal chemistry needs via ChEMBL molecules fragmentation. The ChEMBL database has been chosen as the best representation of the biologically relevant chemical space. In SynthI-Fragmentation, the algorithm fragments molecules in all possible ways according to the specified list of reactions. It then selects the most optimal fragmentation scheme to maximize the number of synthons corresponding to at least one BB from the user-provided library (in our case PBB from eMolecules library). Parts of the molecules not covered by PBB synthons were broken down to the smallest possible synthons. They can be used as inspiration for the enhancement of PBB collections.

### Synthon quality assessment

According to the "rule of two" (Ro2),<sup>12</sup> good quality BB for medicinal chemistry could be defined as those that typically do not add more than 200 Da in MW, two units of clogP, two H-bond donors, and four H-bond acceptors. Those rules were developed in order to select BBs that most likely will produce drug-like compounds. Therefore, the synthons, as fragments of BB that will be added to the final molecule, were filtered according to this rule, and the number of BB compliant to it was assessed for each BB class

### **Diversity analysis**

Diversity analysis of different types of reagents was also performed in the synthon ISIDA descriptor space. It was done by calculating pairwise Tanimoto distance<sup>29</sup> for all synthons within a selected reagent class, followed by the creation of the frequency plot for each of the diversity values. Note, that different synthons may introduce the same structural fragment to the reaction product but with RCs at different positions. The corresponding synthons will have distinct ISIDA descriptors in spite of being based on the same molecular graph, due to the marked-atom mechanism. Two synthons contributing the same fragment and having the RC at the same position, but of different types (allegedly different reaction mechanisms) have however identical ISIDA descriptors (they capture the label position, not its actual value). Such synthons are distinct options covering the same medicinal chemistry need – their existence is practically important because they allow for alternative synthetic pathways, but they are indeed redundant from a structural point of view.

### **GTM**

In chemoinformatics, chemical space can be defined by the N-dimensional molecular descriptor vector, where N is typically very large ( $10^2$ - $10^4$ ) for vectors designed to capture significant chemical information. The most intuitive way to analyze such a complex space is to reduce its dimensionality to obtain a 2D map. Generative topographic mapping (GTM), first proposed by Bishop in 1998,<sup>22</sup> appears as one of the most efficient dimensionality reduction methods.<sup>30</sup> It performs non-linear projections of compounds from the initial multidimensional descriptor space to a 2D latent space - a manifold defined by a set of radial basis functions (RBF). The shape and position of each point of the manifold in the N-dimensional space are determined during its training



– unsupervised fitting to the "frameset" items - molecules used to probe the chemical space of interest. Afterward, the manifold is unfolded back to the planar form – square grid 2D map.

Once trained, the manifold can host not only compounds of the "frameset" but also any external molecules, under the condition that in the multidimensional space, they are residing close to the manifold (log Likelihood applicability domain of GTM<sup>31</sup>). The distinctive feature and the main advantage of GTM is its probabilistic nature, ensured by RBFs. In GTM, molecules are not assigned to a particular point on the map. Instead, each molecule is fuzzily projected over the whole map with larger probabilities ("responsibilities") for nodes situated closer to this compound in the initial space. Such smooth projection enables the creation of GTM landscapes – 2D plots of cumulated responsibilities, colored by average values of different properties, e. g. density, biological activity, physicochemical property, assigned class, etc. One manifold can host multiple landscapes allowing the analysis of multiple libraries according to different properties. It can also be used as a basis for building QSAR models.<sup>24, 30-32</sup>

### **Universal map of synthons (synthons-uMap)**

The "universal" map of synthons (synthons-uMap) is the GTM that would simultaneously host different types of synthons (electrophiles, nucleophiles, radicals, etc.). It can be constructed by optimizing map performance in class separation for the different types of RCs present in synthons.

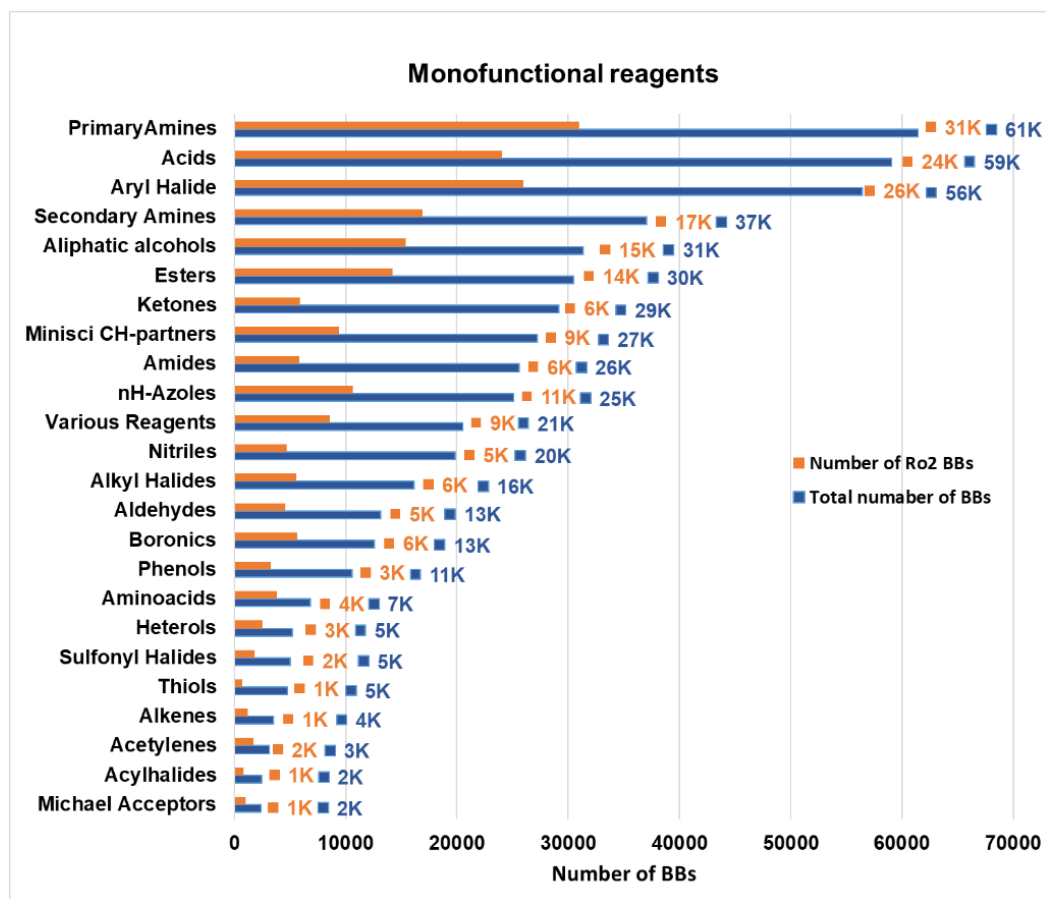
A fixed frame set of 15,255 randomly selected synthons has been used. It contained an approximately equal ratio of synthons obtained by eMolecules in-stock BB library synthonization and ChEMBL fragmentation in order to span the chemical space of both PBB and biologically relevant BB. Seven scoring sets, 15,000 synthons each, were used to evaluate map performance in class separation for electrophiles, nucleophiles, bivalent nucleophiles, bivalent electrophiles, neutral biradicals, electrophilic radicals, and boronic-derived nucleophiles (for Chan-Lam reaction<sup>33</sup> and other couplings). The map was optimized in exploring its (hyper)parameter space by an evolutionary procedure (genetic algorithm or GA) that is customarily employed to tune GTMs,<sup>23, 34, 35</sup> following a Pareto-front-driven multiobjective strategy. This approach considered performances (balanced accuracies) of pairwise classification between 7 synthons classes - monovalent electrophiles, monovalent nucleophiles, boronics-derived nucleophiles, bivalent nucleophiles, bivalent neutral, bivalent electrophiles, radical electrophiles - as independent

objectives. The Pareto front of non-dominated maps was considered the "best" solution (defining the pool of selected individuals allowed producing offspring in the evolutionary strategy).

## RESULTS AND DISCUSSION

391,378 reagents out of 489,781 BB from eMolecules library have been classified and synthonized. The remaining non-classified reagents are either used in heterocyclization reactions that are out of the scope of this analysis or contain conflicting or competing functionalities disqualifying them for combinatorial chemistry.

As a result, 798,643 synthons were generated. In *Figure 1*, one can see the detailed analysis of the availability of monofunctional reagents on the market. The expected leaders of the distribution are amines, acids, and aryl halides. Their "excessive" availability can be explained by the broader usage of combinatorial reactions that employ these reagents. Among all classes of compounds, approximately half of them pass the Ro2 and thus represent the means for drug-like libraries synthesis.



**Figure 1.** Monofunctional commercially available reagents: total number and number of high-quality Ro2 compliant reagents.

## Amines

Despite the strong development of modern organic synthesis, medicinal chemists traditionally use only a tiny fraction of the available reactions, especially in compound library and analogs synthesis.<sup>36</sup> The general criteria for the ideal medicinal chemistry reactions were formulated in 2010 by medicinal chemists at GlaxoSmithKline (GSK)<sup>37</sup> and have not changed significantly over the last decade.<sup>1, 38</sup> Among them, there are requirements for reproducible chemical transformations, applicable to structurally diverse substrates, tolerance for the range of functionalities, simple equipment, and reasonable cost. The reactions that fulfill these criteria, such as amides and sulfonamides formation, alkylations (including reductive amination),  $S_NAr$ /Buchwald and  $C_{Ar}-C_{Ar}$  Suzuki couplings, will always be attractive to medicinal chemists. The majority of such reactions use primary and secondary amines as coupling partners, which explain their leading position on the market.

For more detailed analysis, primary amines have been split into several groups depending on the position of the functional group – aliphatic, benzylic, heterobenzilic amines, anilines, and hetero-anilines. Secondary amines, however, can have even more different topologies, including cyclic and acyclic aliphatic, benzylic and heterobenzylic, and polycyclic amines (**Figure S1**). In both cases, aliphatic amines (cyclic and acyclic) are the most popular (**Figure S2**), which can be explained with current medicinal chemistry demand for the high Fsp3 compounds.<sup>39, 40</sup> Next are the derivatives of hetero-anilines and anilines, which allow the one-stage introduction of new aromatic cycles.

### **Carboxylic acids**

The second place on the market is taken by acids – the main coupling partners of amines. A recent study from AZ indicates that amide couplings sum up to one-third of all the reactions in their electronic notebooks.<sup>41</sup> Similar to purchasable amines, among carboxylic acids, the aliphatic counterparts are dominant (**Figure S3**). They are followed by heteroaromatic and benzoic acids. It should be noted that the homologs of heteroaromatic and benzoic acid - corresponding (hetero)aryl acetic acids - are significantly less present (from 7 to 10 times). It goes in accordance with the observation of AZ made in 2011, stating that synthetically this type of acids is much less accessible.<sup>42</sup> Indeed, out of 148 compounds proposed for the synthesis in AZ, only 17 were successfully made.

### **Arylation reagents**

The leading position of the aryl halides can be explained by the active development of the Pd-mediated Csp2-Csp2 and N-Csp2 couplings.<sup>43</sup> According to Boström's analysis, the Suzuki Csp2-Csp2 coupling is the second most popular transformation after the amide bond formation.<sup>36</sup> The same was later confirmed by Eli Lilly's analysis of the reactions performed using their ASL robotic synthesis system<sup>44</sup> and AbbVie's high-throughput chemistry department.<sup>45</sup> The high reproducibility of Csp2-Csp2 coupling together with its modern improvement made this reaction suitable for automation. In 2015 Burke designed a generalized automated process for the C-C couplings by analogy with well-known automated peptide synthesis based on amide bond creation.<sup>46</sup> Despite such outstanding achievements in Suzuki couplings development, the

commercial accessibility of organoboron building blocks is still significantly lower than (hetero)aromatic electrophiles (**Figure 1**).

Buchwald-Hartwig (BH) amination is also very popular. The power of this reaction lies in the ability to couple two fragments with minimal addition of rotatable bonds in the final structure. However, its success rate still hardly exceeds 45% due to the lack of a general catalytic system for diverse substrates. Besides, the reactivity in BH amination for the significant portion of available amines has not been experimentally validated yet and is hard to predict. At the same time, the active development of high-throughput experimentation (HTE) chemistry,<sup>47, 48</sup> and machine learning approaches<sup>49</sup> significantly accelerates the identification of effective catalytic systems and the scope of their application.

The alternative well-studied metal-free transition - "classical"  $\text{S}_{\text{N}}\text{Ar}$  amination cannot compete with the BH reaction. It appears that among all aryl halides, only a limited fraction bears activated halogen atoms suitable for non-catalytic amination (**Figure S4**). Interestingly, in the case of (hetero)aromatic chlorides, almost all of them (13,305 out of 14,697) bear activated chlorine atoms likely to undergo  $\text{S}_{\text{N}}\text{Ar}$  reactions. It could be explained by the fact that early conditions for the Suzuki coupling were inapplicable for the aromatic chlorides. However, the opposite situation is observed for aromatic bromides, which are convenient partners for the Suzuki couplings. Indeed, only 10% of aryl bromides are suitable for metal-free amination (3,664 out of 34,586). The number of bromides for the  $\text{S}_{\text{N}}\text{Ar}$  reaction is comparable with hetero(aromatic) compounds bearing an active fluorine atom (3,361), but the number of identified  $\text{S}_{\text{N}}\text{Ar}$  iodides (957 of 6,330) is significantly smaller.

### Alkylation agents

The  $\text{C}(\text{sp}^3)\text{-N}$  bond creation is also very popular and accounts for up to 10.6% of all reactions performed in industrial medicinal chemistry departments, according to Vernalis statistics<sup>50</sup>. The alkylation or reductive amination is regularly used for that aim. Among these two reactions, the reductive amination is slightly more preferable,<sup>45, 51</sup> because it is more selective and allows avoiding a significant number of by-products observed during alkylation. Nevertheless, this approach has its limitations, caused by the low diversity of the commercial carbonyl compounds. In the case of aldehydes, the most popular reagents are aromatic (6,178) and heteroaromatic ones

(5,064), generating benzylic-type synthons (**Figure S5**). Aliphatic aldehydes are less represented, especially (het)aryl acetic ones (only 129 BB), due to their extremely low stability and high rate of self-condensation. Ketones are better represented in commercial catalogs, but there is still a lack of the most interesting for medicinal chemists cyclic ketones (only 7,197 from which only 2,447 pass Ro2).

Expanding the space of the synthons for alkylation could be achieved by commercially available alkyl halides. Commercially available alkyl chlorides and bromides are preferred over iodides and primary alkyl halides are significantly more accessible than secondary ones (**Figure S6**). (Hetero)benzylic primary alkyl halides (4,305) are less represented in comparison with the corresponding aldehydes (11,242). An even higher difference is observed while comparing secondary halides (2,445 in total) and ketones (29,152). It can be explained by the lower shelf-life time of alkyl halides. Indeed, many of them are obtained from corresponding alcohols prior to synthesis. Moreover, nowadays, efficient methods for the in situ alkylating agent generation (including chlorides, bromides, and iodides) were developed. For example, recently, SO<sub>2</sub>F<sub>2</sub>-mediated in situ generations of 1° and 2° alkyl halides were proposed.<sup>52</sup>

Other very efficient alkylating reagents - sulfonate esters, like mesylates, tosylate, and triflate - also have low shelf-life time. It makes their precursors, alcohols, more attractive for purchase and storage as latent alkylators. There are also ongoing attempts to develop direct methods for the alkylation of amines with alcohols. Among them, the advanced reaction conditions for the well-known Mitsunobu reaction<sup>53</sup> that allows basic amine usage<sup>54</sup> and a novel Ru-based catalyst system for hydrogen borrowing reaction, proposed by GSK in 2009.<sup>55</sup> Therefore, it is not surprising that representation of this reagent class on the market is comparable with secondary amines. In contrast to alkyl halides, there is approximately the same number of primary and secondary alcohols with a slight excess of the latter, while the number of benzylic and heterobenzylic alcohols is comparable to corresponding alkyl chlorides (**Figure S7**).

### **Sulfur-containing BB**

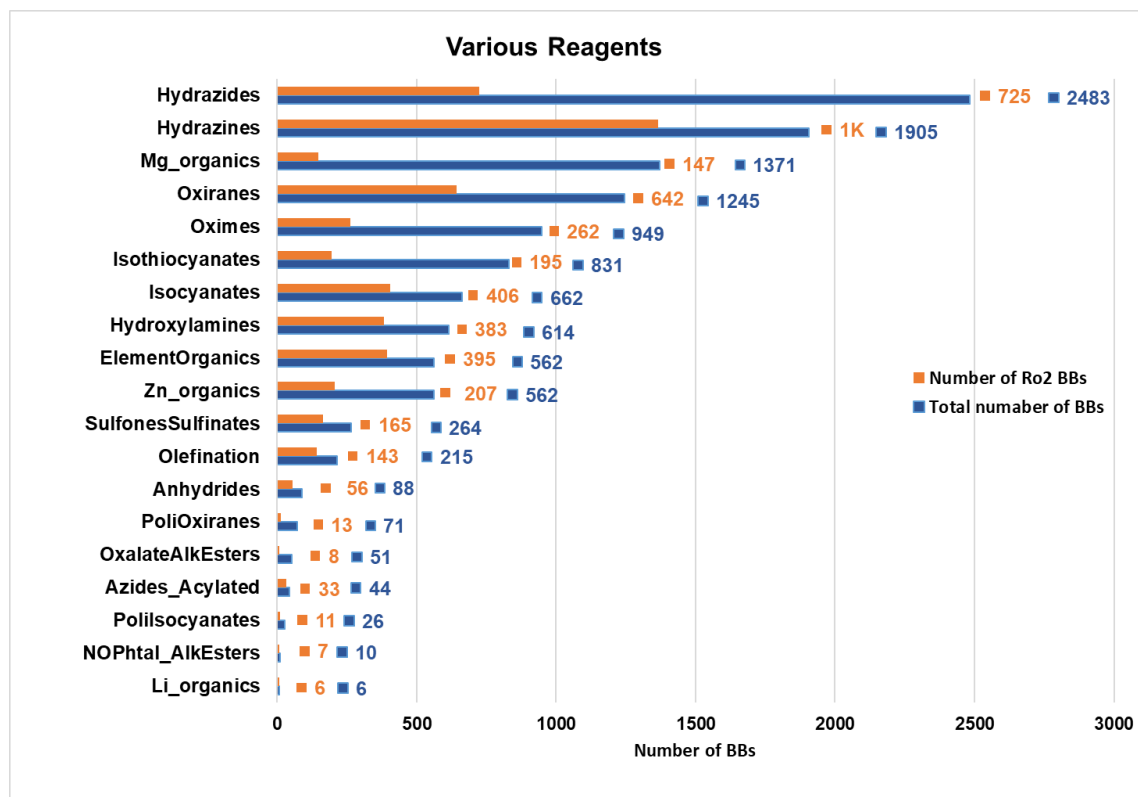
Surprisingly, despite the high popularity of sulfonamides in medicinal chemistry, the number of available sulfonyl chlorides (4,099) and especially fluorides (582) is relatively low (**Figure S8**). The leading position among them is taken by aryl sulfonyl chlorides (2,079), which can be

explained by their higher stability in long-term storage in comparison to alkyl and heteroaryl sulfonyl chlorides, that can undergo SO<sub>2</sub> extrusion. The seminal paper addressing this stability issue was published by Pfizer in 2006.<sup>56</sup> It was also shown that sulfonyl fluorides could become a convenient replacement for sulfonyl chlorides, as they are more thermodynamically stable, resistant to reduction, and chemoselective towards sulfonylation products. However, such an approach has not gained attention until the introduction of Sulfur(VI) Fluoride Exchange (SuFEx) reaction for click chemistry by Sharpless et al. in 2014.<sup>57</sup> Since then, numerous works have been published on synthesis and usage of SuFEx building blocks.<sup>58-62</sup> However, as one can see in **Figure S8**, the market did not have enough time to react to the newly emerged tendency, and there is only a limited number of such reagents available yet.

Another unexpected observation is that the total number of thiols on the market is rather low (4,788), even though S-alkylation is one of the most well-studied reactions in combinatorial chemistry. It can be explained by the complicated storage conditions required for these reagents. Since thiols can easily undergo oxidation and form disulfides, they should be stored in ampules with an inert atmosphere. The heteroaromatic thiols are the most populated group (3,688) due to their additional stability gained via thione-thiol tautomerism.

### Other reagents

A recent statistic published by Pfizer indirectly proves the above-mentioned tendency in the late-stage combinatorial reactions popularity. In the course of its Quick Building Blocks program, out of all BB, they have used 29% of acids amine – 21%; alcohol - 9%; aryl halide - 9%; mono-BOC diamine - 6% ; aniline - 5%; aldehyde - 4%; aryl boronic acid - 4% and sulfonyl chlorides only 3%.<sup>63</sup> At the same time, there are also less represented classes of reactions and reagents that are widely used for larger BB synthesis in the early stages of the synthetic pathway. In **Figure 2**, one can see that among various reagent classes, the most numerous are hydrazides and hydrazines. Iso(thio)cyanates, hydroxylamines, and element-organics occupy the middle position. Among metallorganics, Grignard reagents expectedly are the most numerous class. Organozinc BB account for two times fewer compounds, and there are only 6 Li-containing reagents.



**Figure 2.** Commercially available various reagents: total number and number of high-quality Ro2 compliant BB.

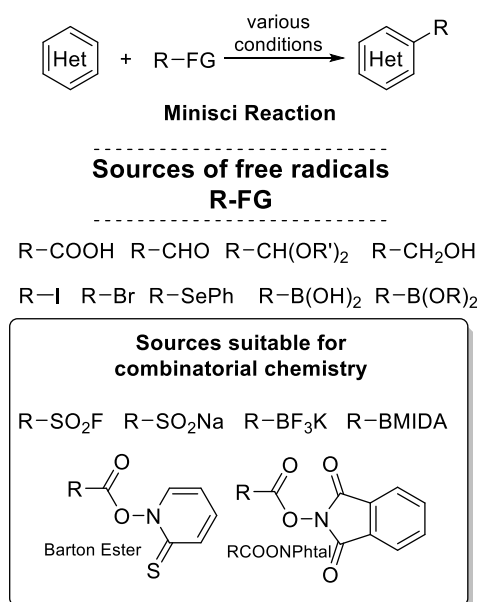
### Reagents for novel reactions

Some novel synthetic methodologies became more frequently used by medicinal chemistry. This was influenced by the emergence of new methodologies, instruments, and techniques for elaborating and carrying out combinatorial reactions. Among them, we can highlight the automatic optimization of the reaction conditions at nanomolar scale,<sup>47</sup> a robustness screen that allows quick determination of the scope of the application of substrates with additional functional groups,<sup>64</sup> new selective radical processes,<sup>65</sup> photoredox catalysis,<sup>66</sup> a new generation of click chemistry,<sup>67, 68</sup> automated interactive cross-coupling,<sup>46</sup> and late-stage functionalization.<sup>69</sup> However, the above-mentioned distribution of BB hinders the development of such novel combinatorial reactions due to the poor representation of necessary reagents among the commercially available. One such example is a Minisci reaction, which efficacy is based on an effective in-situ generation of free radicals.<sup>70</sup>

As pointed out in **Figure 3**, many types of reagents can be used as a source of free radicals. Nevertheless, only some of them, such as R-SO<sub>2</sub>F, R-SO<sub>2</sub>H salts, RCOONPhtal, R-BF<sub>3</sub>K and R-



BMIDA are useful for the combinatorial synthesis of compound libraries, which could fill the "white spots" in chemical space. So, despite sufficiently numerous (27,204 BB) 'Minisci\_CHpartners' (**Figure 1**), the implementation of this reaction in a combinatorial manner is limited by the scarcity of coupling partners – there are only 716 R-BF<sub>3</sub>K<sup>+</sup> and R-BMIDA (418 of them are Ro2 compliant) and even less R-SO<sub>2</sub>F (582) and RCOONPhtal esters (only 10). Thereby chemical companies could focus their attention on this problem and direct their efforts towards BB catalogs enhancement with such needful reagents.



**Figure 3.** Representative types of reagents for Minisci Reaction.

### Polyfunctional BB

Apart from the monofunctional reagents, appropriately protected bi- and trifunctional BB are required for optimal combinatorial library construction. Among the bifunctional ones, the absolute leaders are different derivatives of amino acids (**Figure 4**) due to the extreme popularity and automation of peptide synthesis. Other large classes are Boc-protected diamines and functional aryl halides. Such distribution reflects the same tendencies that have been observed and explained for monofunctional building blocks. Polyfunctional reagents play the role of molecular cores around which a diverse set of monofunctional partners allows the creation of large combinatorial libraries. Therefore, bi- and especially trifunctional BB are crucial for the synthesis of DNA-encoded libraries (DEL). Thus their availability is affected by the popularity and efficiency of the

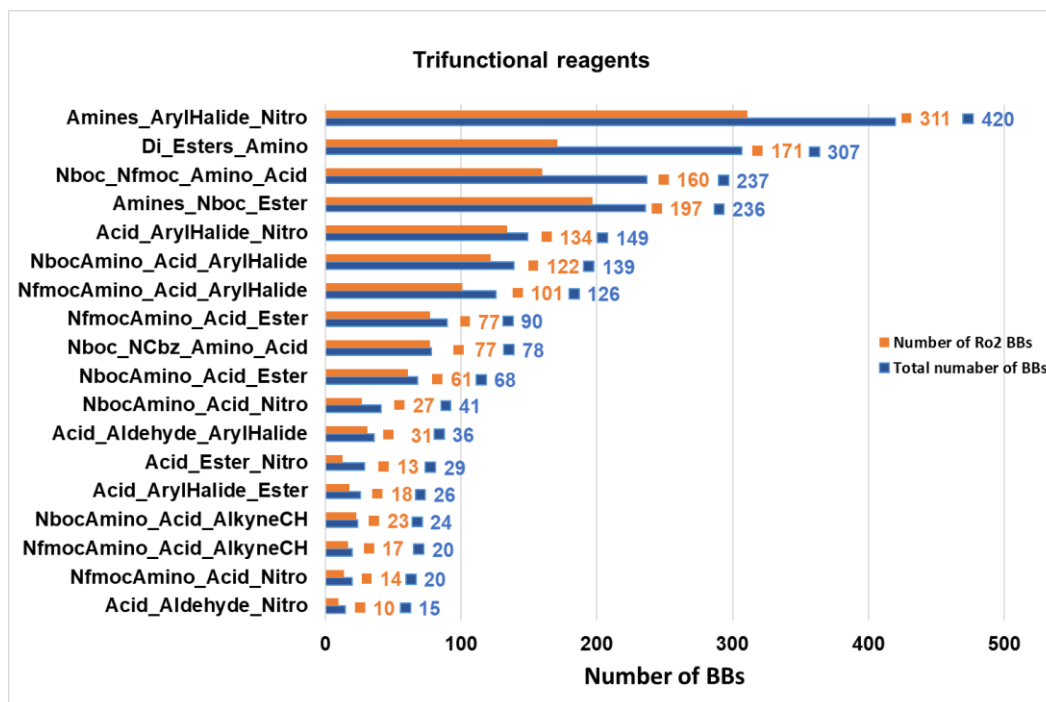
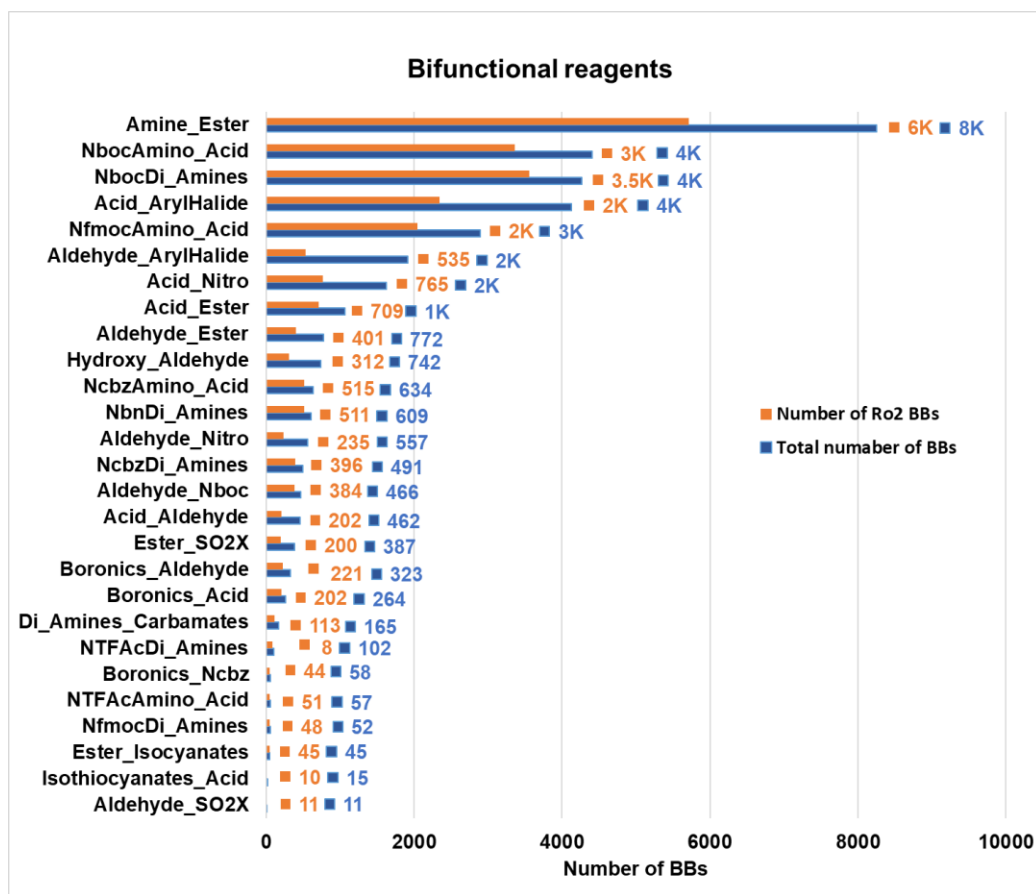
reactions adapted for this technology. Considering the relatively recent development of DEL, a limited number of corresponding reagents on the market is understandable.

### **Medicinal chemists' highlights**

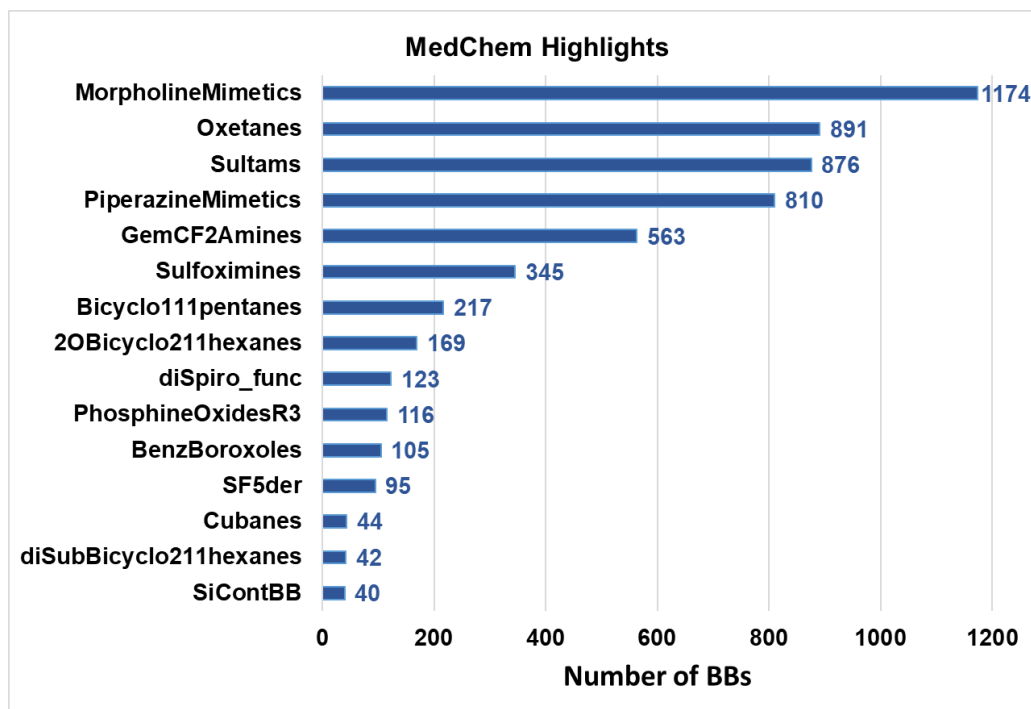
Earlier in this analysis, the main focus was set on functional group types that define the BB that may be successfully used in a reaction. However, what is even more important for medicinal chemists is what structural moieties will be introduced and how these will influence the pharmacodynamics or pharmacokinetic properties of the synthesized compound. Considered motifs emerge from "breakthrough" approvals of a new drug containing unusual structural moieties. They include morpholine and piperazine bioisosters,<sup>71-73</sup> unusual fluorine-containing aliphatic substituents,<sup>74, 75</sup> sulfoximines,<sup>76</sup> phosphine oxides,<sup>77</sup> silicon-containing isosteres<sup>78</sup> and non-classical sp<sup>3</sup>-enriched benzene isosteres, such as bicyclo[1.1.1] pentanes, cubanes, etc.<sup>79</sup> In **Figure 5**, one can see that there are only a limited number of BBs bearing such structural motifs. The distribution leader is morpholine and piperazine mimetics, oxetanes, and sultanes, while there are less than a hundred cubanes, disubstituted bicyclo[2.1.1]hexanes, and silicon-containing BB.

### **The ability of the BB market to face current medicinal chemistry needs**

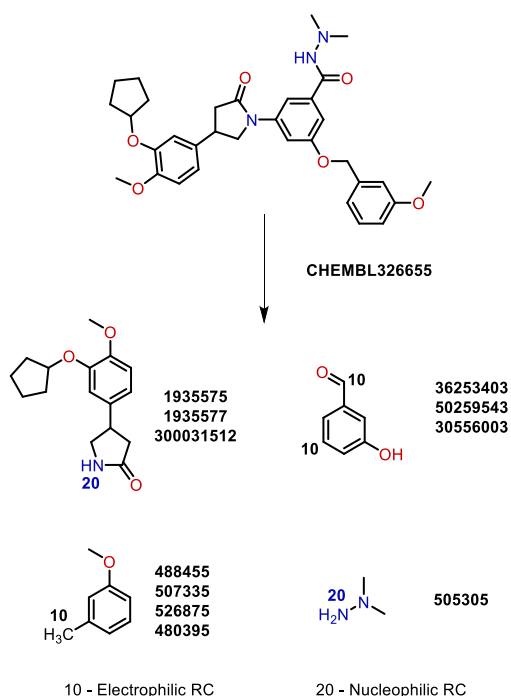
For evaluation of the ability of PBB to face medicinal chemistry needs, ChEMBL library, as a source of biologically relevant compounds, was fragmented using SynthI (see example in **Figure 6**). As a result, around 35% of ChEMBL molecules were fragmented into synthons that are all found in the eMolecules library. Around 5% of ChEMBL was not cut at all due to the small size of the molecules and lack of synthetically accessible acyclic bonds (heterocyclization was not taken into account). The remaining 60% of compounds have some but not all of synthons available – they include at least one synthon out of the scope of the eMolecules library.



**Figure 4.** Polyfunctional commercially available reagents: total number and number of high-quality Ro2 compliant molecules.



**Figure 5.** Commercially available reagents, containing highly attractive structural motifs.



**Figure 6.** Example of ChEMBL molecule fragmentation towards commercially available synthons (eMolecules identifiers of corresponding BB are provided).

For a more detailed analysis, in addition to the groups of synthons explained in *Table S1*, electrophiles were further subdivided into acylating and sulfonylation agents, C-alkyl and C-aryl

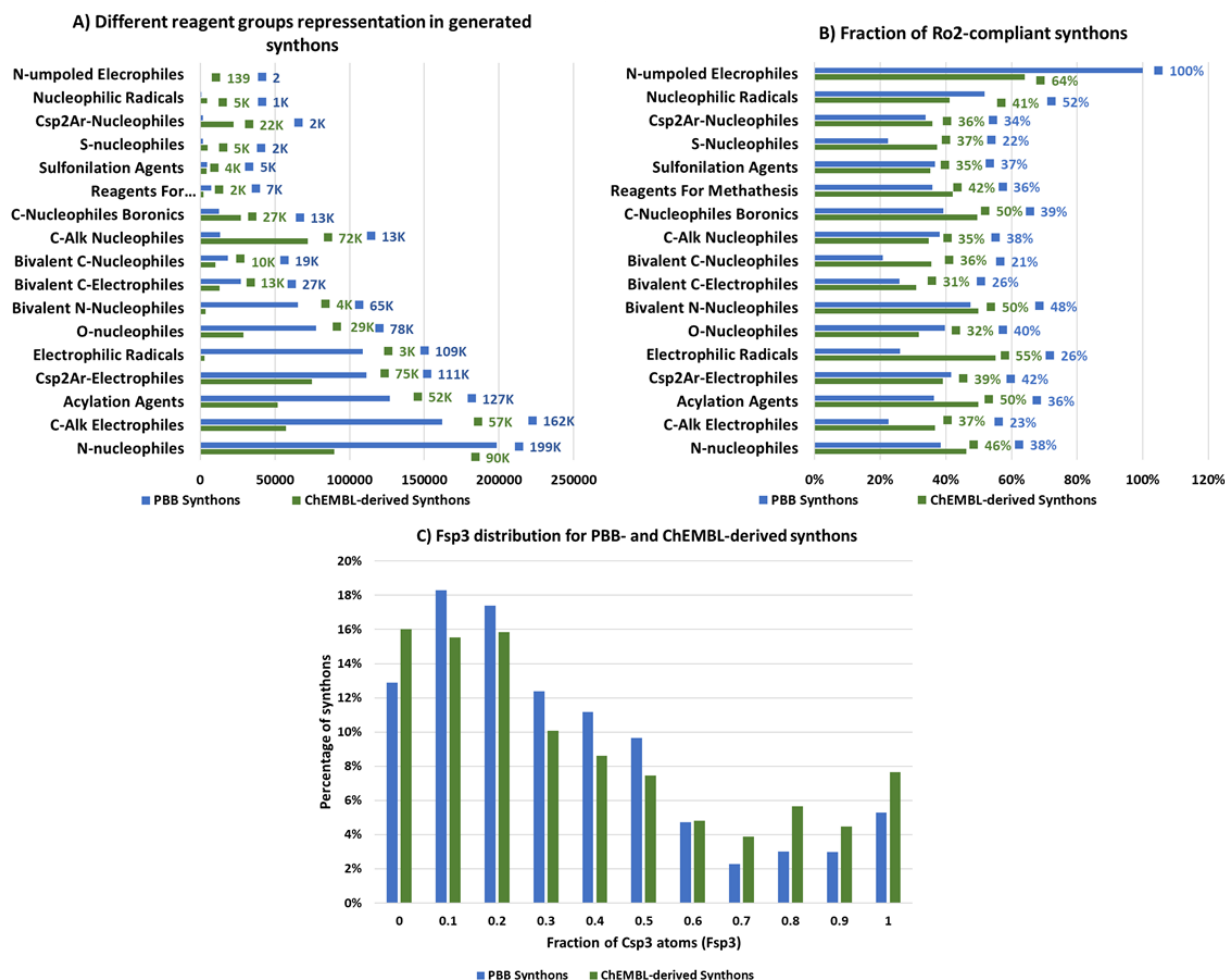
electrophiles. The nucleophiles were split into N-, O-, S-, C-alkyl, and C-aryl nucleophiles. The populations of all synthon groups have been analyzed in **Figure 7 (A)**, the fraction of Ro2-compliant synthons – in **Figure 7 (B)** and Fsp3 distribution is given in **Figure 7 (C)**. In addition, the relative diversity distribution for bivalent and monovalent synthons classes is provided in **Figure S9** and **Figure S10** respectively.

In comparison with synthons generated from ChEMBL, the chemical market offers an abundance of reagents producing N-, O-nucleophiles, classical electrophiles, bivalent synthons, and electrophilic radicals. At the same time, there are several underrepresented synthons classes: all types of C-nucleophiles (Csp3-, Csp2- and C-boronics), S-nucleophiles, nucleophilic radical, and N-electrophiles. It goes in correspondence with conclusions derived in the previous chapter. However, synthons diversity for all the groups is higher for corresponding ChEMBL-specific synthons subsets (**Figure S9** and **Figure S10**), especially in the case of bivalent nucleophiles and electrophiles. This fact is well expected as the decision to add the BB to the purchasable collection is influenced by many practical factors (discussed earlier in this paper), while many ChEMBL compounds were synthesized by academic laboratories that are not limited by such factors and are more inclined towards complicated and challenging synthetic procedures.

As one can see from **Figure 7 (B)**, the fractions of Ro2-compliant synthons in PBB- and ChEMBL-derived collections are comparable for almost all synthons classes and usually do not exceed 50%. The only exception is the less populated class – N-unpoled electrophiles. The eMolecules collection contains only two respective reagents which both are Ro2 compliant. The low number of Ro2-compliant synthons in ChEMBL can be explained by the fact that ChEMBL fragmentation was biased towards PBB. Thus, if large ChEMBL-derived synthon is present in the eMolecules library, it is never fragmented even if some bonds can be cleaved by SynthI. The synthon's availability and the smallest number of reaction steps are the main priority. The synthons are fragmented to the smallest possible pieces only if they are not present in PBB. Note, that three ChEMBL-derived synthons classes - electrophilic radicals, acylation agents and N-nucleophiles - have higher content of Ro2-compliant structures than PBB **Figure 7 (B)**. However, all of them are significantly better represented in PBB **Figure 7 (A)**.

The fraction of carbon atoms with sp<sup>3</sup>-hybridization (Fsp3) allows quantitative analysis of the “flatness” of given compounds (in our case synthons). **Figure 7 (C)** demonstrates that the majority of

synthons exhibit a low value of Fsp3 – around 81% of PBB- and 73% of ChEMBL-derived synthons have Fsp3 lower than 0.5. The shift towards higher values observed for the ChEMBL-derived synthons signifies the need for more Csp3-enriched BBs in the commercial collections.



**Figure 7.** Comparison of PBB- and ChEMBL-derived synthons: A) popularity of different types of synthons; B) fraction of Ro2-compliant synthons; C) Fsp3 distribution.

## GTM-based analysis of synthons

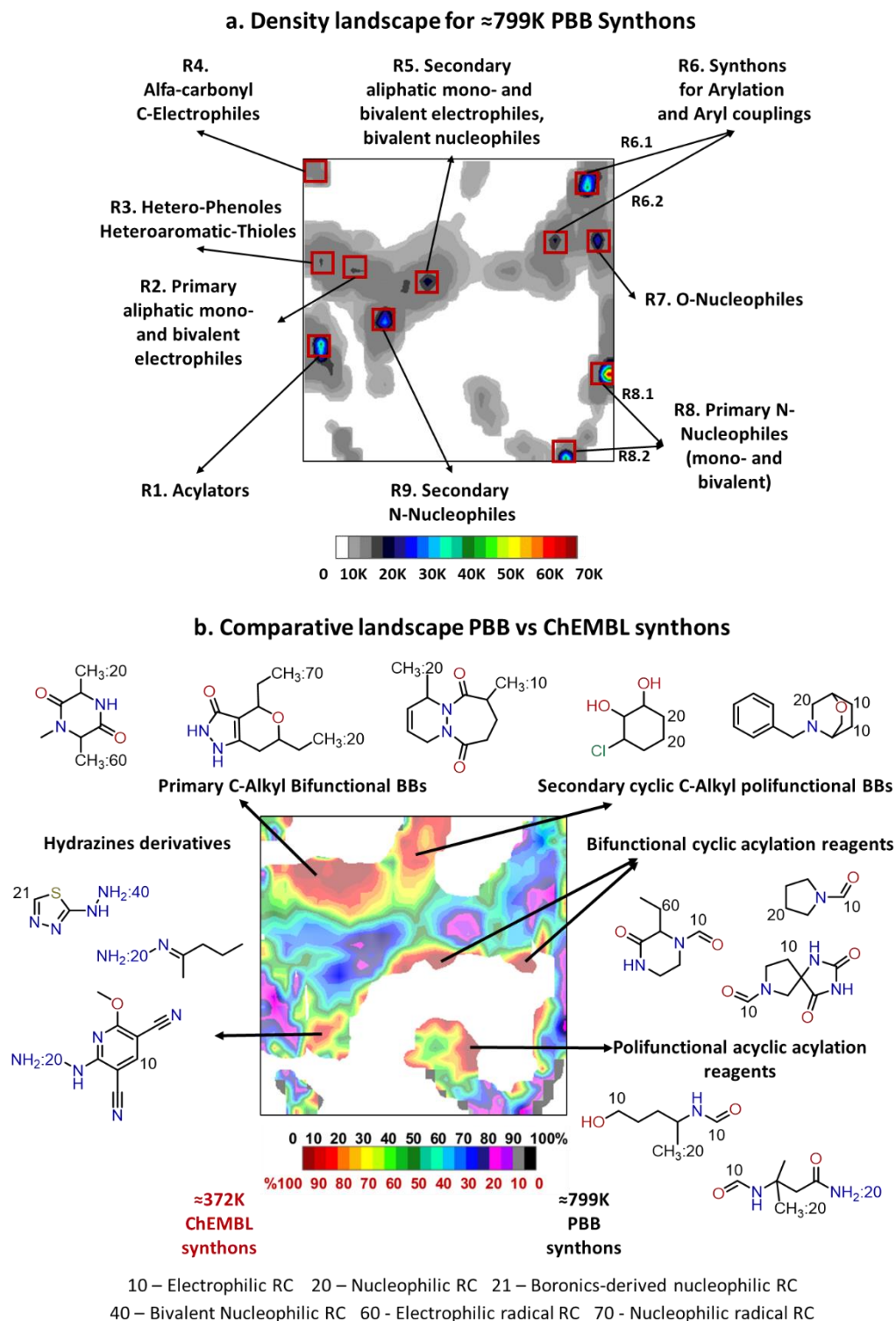
As a result of GA Pareto optimization, a synthons-uMap was selected out of thousands of evaluated options. This map is based on the atom-centered fragments of 1-2 atoms radius, including atoms and bond information. These descriptors are highly sensitive to RC position, which allows distinguishing between synthons with different reactivity due to the inductive, mesomeric, or steric effects. The manifold consists of a grid of 29\*29 nodes coupled with 25\*25 RBFs. This map

provides synthon class separation with average balanced accuracy - BA of 0.9 (the lowest BA=0.79 for separation of C-nucleophiles from all other classes).

In **Figure 8(a)**, one can see the density distribution for PBB-based synthons. The color code reflects the number of synthons in each point of the map – grey regions correspond to the minimally populated areas of the chemical space, while multicolored ones depict high-density picks. In agreement with previous synthon population analysis, the highest density is observed in the south-eastern part of the map. It corresponds to the primary N-monovalent and -bivalent nucleophiles produced by aliphatic amines and anilines (R8.1). Interestingly, primary hetero anilines form a separate cluster of slightly lower density further on the south (R8.2). At the same time, secondary N-nucleophiles are situated quite far from the primary ones in the central part of the map (R9). They are surrounded by acylation agents (R1) from one side and secondary aliphatic synthons with RC on the carbon atom from the other – mono- and bivalent C-electrophiles and bivalent C-nucleophilic synthons (R5). This is expected, as the ISIDA descriptors are sensitive to the position of the RC (marked atom) but not to the actual value of the atom label (encoding the type of intermediate). Therefore, C-electrophiles and C-nucleophiles (mono- or bivalent alike), can be found in the same region, but secondary (R5) and primary (R2) aliphatic synthons with RC on carbon atom are spatially separated. So are aliphatic (R2, R5) and aromatic (R6) synthons.

Similar to the primary N-nucleophiles in the regions R8.2 and R8.3, arylation agents (Csp<sup>2</sup>Ar-electrophiles, electrophilic radicals, Minisci CH-partners, aryl-boronics etc.) are split into two clusters with high density. The more crowded area is dominated by phenyl and  $\alpha$ -pyridine synthons (R6.1). At the same time, the region with relatively moderate occupancy is populated by  $\gamma$ -heteroaryl synthons, usually with a higher number of heteroatoms (R6.2). The latter is neighboring the region R7, occupied by O-nucleophiles – aliphatic, benzylic alcohols, and phenols. Meanwhile, hetero-phenols and heteroaromatic thiols populate the area on the opposite part of the map (R3).

The fact that in the same zone of the map one can find structurally similar or even identical synthons differing only in the nature of the RC is actually an advantage. The map can thus be used to search alternative synthesis ways in situations where the same structural moiety can be provided by building blocks of radically different reactivity, applicable in distinct synthetic paths. For example, bivalent C-electrophiles and C-nucleophiles, intermediates in Knoevenagel-, Wittig-,



**Figure 8.** GTM analysis of synthons: a) density distribution of PBB synthons (color code reflects number of compounds in each point of the map); b) comparison of PBB synthons (black areas) and ChEMBL-derived synthons (red regions). ChEMBL-specific regions are profiled with examples of respective synthons.



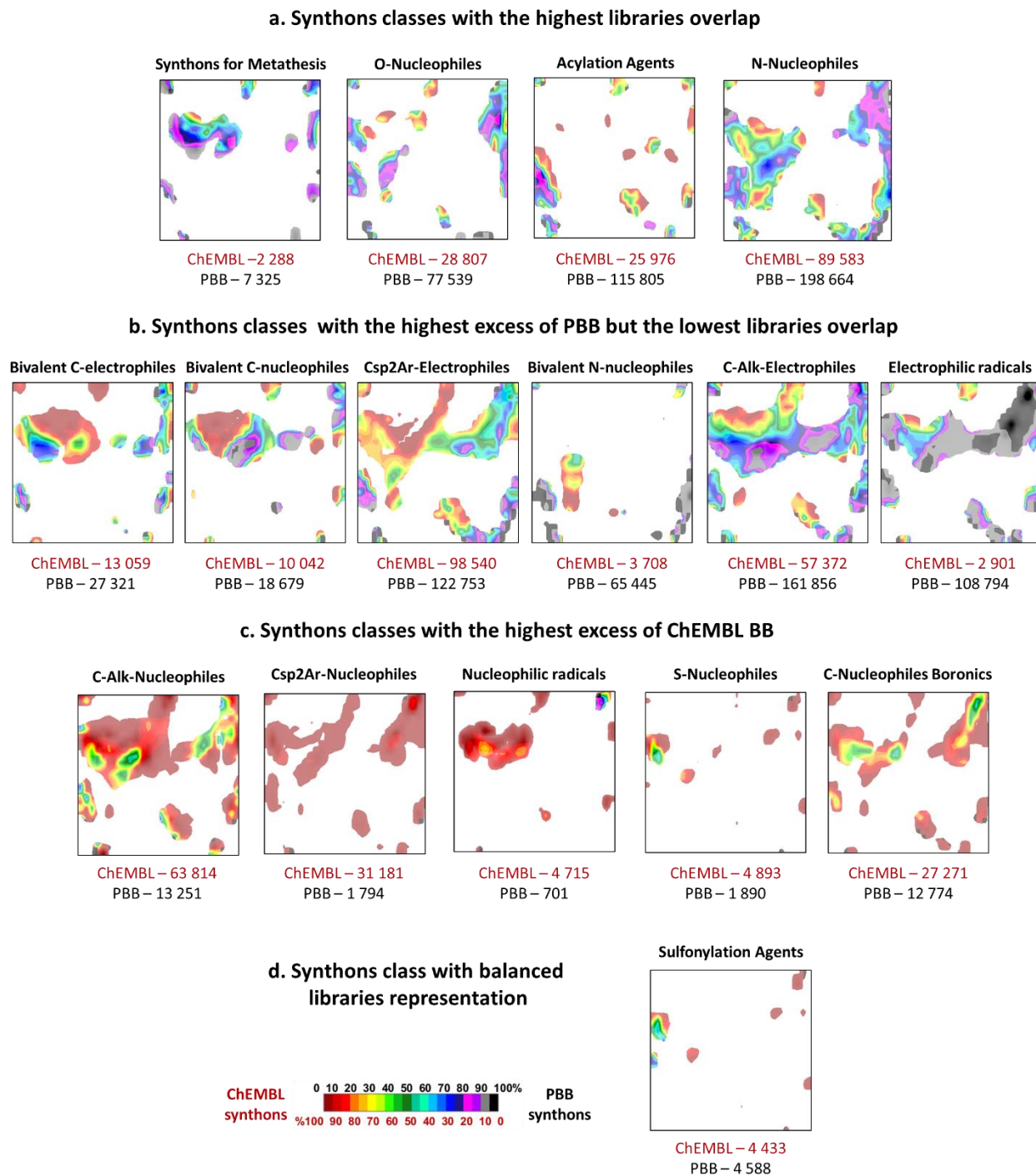
Julia-Kocienski- types of olefination occupy the same areas as reagents for metathesis – another reaction for double C=C bond formation.

**Figure 8(b)** shows the comparative landscape featuring PBB synthons (black color) versus synthons obtained via ChEMBL fragmentation (red color). All colors in between correspond to the mixed regions of different compositions (see the scale). It appears that even though the number of PBB synthons is more than two times higher than the number of ChEMBL-derived synthons, there are still large ChEMBL-specific areas of the chemical space (red regions). These zones mostly correspond to the polyvalent synthons, which, as has been discussed earlier, are underrepresented on the market. In addition, the majority of synthons residing in ChEMBL-specific regions contain heterocycles, but heterocyclization processes were excluded from this analysis. Notice that Ro2 filtering of synthons almost doesn't change the shape of the comparative ChEMBL/PBB landscape (see **Figure S11** in SI). This means that Ro2-compliant synthons cover the prevailing majority of the whole synthon's chemical space.

In order to obtain a better understanding of the chemical space of different synthons classes, 16 comparative ChEMBL vs PBB landscapes for each group analyzed above were constructed (**Figure 9**). Their comparison shows that despite lower diversity (**Figure S9** and **Figure S10**) of PBB synthons in all categories, there are still four classes that largely cover the chemical space of the respective ChEMBL synthons (**Figure 9 (a)**). Among them, there are synthons for metathesis, O- and N-nucleophiles, and acylation agents. In all these classes, there is a significant abundance of PBB synthons over ChEMBL-derived ones.

However, the high number of synthons does not always guarantee better coverage of biologically relevant synthons space. Indeed, bivalent electrophiles and nucleophiles, C-electrophiles, and electrophilic radicals are also more numerous within the PBB synthons, but the overlap between commercially available and biologically relevant synthons is the smallest for these subsets (**Figure 9 (b)**). There are large areas exclusively occupied by representatives of only one library, which means that the abundance of such synthons on the market still leaves room for improvement of the quality and structural diversity of corresponding BB. It mostly concerns areas associated with polyfunctional synthons containing more than one RC (**Figure 8(b)**).

The trends outlined in **Figure 7** are clearly seen in the comparative landscapes – there is a significantly higher portion of red areas for C- and S-nucleophilic synthons (**Figure 9 (c)**). Interestingly, even in the case of equivalently represented PBB and ChEMBL sulfonylation agents, there are still areas of biologically relevant synthons space not covered by PBB.



**Figure 9.** Synthons classes-based comparison of PBB synthons (black areas) and ChEMBL-derived synthons (red regions).

## CONCLUSIONS

In this work, commercially available BB, provided by eMolecules, were analyzed in terms of purchasability, quality, diversity, and ability to face current medicinal chemistry needs. The last was achieved by fragmenting biologically relevant molecules from the ChEMBL database with the help of SynthI – a knowledge-based reaction toolkit for library design and analysis. The resulting synthons were compared to those generated from PBB. This led to a detailed, comprehensive analysis of PBB in a medicinal chemistry context.

It was shown that the most represented classes of BB – amines, acids, aryl halides, and aliphatic alcohols – mirror the popularity of the respective reactions – amide formation, Pd-mediated couplings, Buchwald-Hartwig amination, alkylation etc. However, the existence of well-studied reactions is not the only factor defining reagent availability on the market. Indeed, sulfonate esters, secondary and (hetero)benzylic primary alkyl halides are far less present than other alkylation agents - alcohols, ketones, and aldehydes, respectively, due to their lower shelf-life time. The low number of S-nucleophiles can be explained by complicated storage conditions, while the lack of SuFEx reagents and polyfunctional BB – by the relative youth of the efficient methodologies involving these reagents.

It was also noted that the reported distribution of BB could limit the development of novel combinatorial techniques (nanomolar scale, robustness screen, photoredox catalysis, new generation of click chemistry, automated interactive cross-coupling, and late-stage functionalization). These are disfavored by the poor representation of necessary reagents, e.g., R-SO<sub>2</sub>F, R-SO<sub>2</sub>H salts, RCOONPhtal, R-BF<sub>3</sub>K, and R-BMIDA, SuFEx, and polyfunctional BB for DEL design.

Comparison of PBB- with ChEMBL-derived synthons reveals that the internal diversity among members of the same synthons class is significantly better for ChEMBL-derived synthons. It was shown that there is a lack of C- and S-nucleophiles and nucleophilic radicals, while O- and N-nucleophiles and electrophilic reagents are overrepresented on the market.

Synthon-uMap, constructed in this work, allows distinguishing structurally equivalent synthons that differ only in the position of the reactive center. The types of reactive center, however, are not differentiated. Therefore, two synthons corresponding to the same structural moiety and having

the reaction center at the same position, but of different type (correspond to different reaction mechanisms) will be projected in the same area of the map. In such a way, synthons-uMap can be used for the search of alternative BBs that as a result of completely different synthetic procedures will still introduce the same structural fragment to the reaction product.

GTM analysis with the help of this map allowed identifying that only in the case of four synthons classes PBB synthons cover largely ChEMBL-derived synthons chemical space: synthons for metathesis, acylation agents, O- and N-nucleophiles. For the other groups, even for those with high PBB synthons excess, there are plenty of ChEMBL-specific areas of chemical space without any purchasable counterparts. Most of these areas correspond to polyfunctional BB which are underrepresented on the market.

All of these findings lead to the conclusion that there are plenty of possibilities for BB libraries improvement – starting with enlargement of underrepresented BB classes subsets and finishing with improving diversity and biological relevance of PBB.

## ASSOCIATED CONTENT

### Supporting Information.

The Supporting Information is available free of charge via the Internet at <http://pubs.acs.org>.

It contains:

**Table S1.** Synthons labels and examples of reagents that generate them.

**Figure S1.** The schematic representation of different topologies for secondary amines.

**Figure S2.** Availability of primary and secondary amines: total number and number of high-quality Ro2 compliant molecules.

**Figure S3.** Commercially available carboxylic acids: total number and number of high-quality Ro2 compliant molecules.

**Figure S4.** Commercially available aryl halides: total number and number of high-quality Ro2 compliant molecules.

**Figure S5.** Commercially available aldehydes: total number and number of high-quality Ro2 compliant molecules.

**Figure S6.** Availability of alkyl: total number and number of high-quality Ro2 compliant molecules.

**Figure S7.** Availability of aliphatic alcohols: total number and number of high-quality Ro2 compliant molecules.

**Figure S8.** Availability of sulfonyl halides: total number and number of high-quality Ro2 compliant molecules.

**Figure S9.** Relative diversity distribution for bivalent synthons classes.

**Figure S10.** Relative diversity distribution for monovalent reagents' classes.

**Figure S11.** Comparative landscapes of unfiltered and Ro2-compliant PBB- and ChEMBL-derived synthons.

### **Data Availability Statement**

The list of ChEMBL-derived synthons generated in this work is publicly available at DOI:10.5281/zenodo.5085657

### **AUTHOR INFORMATION**

#### **Corresponding Author**

Prof. A. Varnek, E-mail: [varnek@unistra.fr](mailto:varnek@unistra.fr).

#### **ORCID**

Yuliana Zabolotna: 0000-0001-9068-612X

Dmitriy M. Volochnyuk: 0000-0001-6519-1467

Dragos Horvath: 0000-0003-0173-5714

Dmitriy M.Volochnyuk: 0000-0001-6519-1467

Sergey V.Ryabukhin: 0000-0003-4281-8268

Konstantin S.Gavrilenko: 0000-0001-7487-6660

Oleksandr Oksiuta: 0000-0003-3049-0373

Yurii S.Moroz: 0000-0001-6073-002X

Gilles Marcou: 0000-0003-1676-6708

Alexandre Varnek: 0000-0003-1886-925X

### Author Contributions

The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

### Funding Sources

YuZ thanks the Doctoral school of the University of Strasbourg for a PhD fellowship

### ACKNOWLEDGEMENTS

The authors are grateful to eMolecules, Inc. for the provided library of commercially available BB, featured in this analysis.

### ABBREVIATIONS

AZ , AstraZeneca; BB, building blocks; BH, Buchwald-Hartwig; DEL , DNA-encoded libraries; GA, genetic algorithm, GSK, GlaxoSmithKline, GTM, Generative Topographic Mapping;PBB, purchasable building blocks; SynthI, Synthons Interpreter; RBF, radial basis functions; Ro2, “rule of two”; SuFEx , Sulfur(VI) Fluoride Exchange; synthons-uMap , "universal" map of synthons chemical space; THE, high-throughput experimentation.

### REFERENCES

1. Grygorenko, O. O.; Volochnyuk, D. M.; Ryabukhin, S. V.; Judd, D. B., The Symbiotic Relationship between Drug Discovery and Organic Chemistry. *Chem. Eur. J.* **2020**, *26*, 1196-1237.
2. Volochnyuk, D. M.; Ryabukhin, S. V.; Moroz, Y. S.; Savych, O.; Chuprina, A.; Horvath, D.; Zabolotna, Y.; Varnek, A.; Judd, D. B., Evolution of Commercially Available Compounds for Hts. *Drug Discov. Today* **2019**, *24*, 390-402.
3. Zabolotna, Y.; Lin, A.; Horvath, D.; Marcou, G.; Volochnyuk, D. M.; Varnek, A., Chemography: Searching for Hidden Treasures. *J. Chem. Inf. Model.* **2021**, *61*, 179-188.
4. Walters, W. P., Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62*, 1116-1124.
5. Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E., Drug-Like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643-51.
6. Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A., Drug- and Lead-Likeness, Target Class, and Molecular Diversity Analysis of 7.9 Million Commercially Available Organic Compounds Provided by 29 Suppliers. *J. Chem. Inf. Model.* **2010**, *50*, 470-9.

7. Lucas, X.; Gruning, B. A.; Bleher, S.; Gunther, S., The Purchasable Chemical Space: A Detailed Picture. *J. Chem. Inf. Model.* **2015**, *55*, 915-24.
8. Petrova, T.; Chuprina, A.; Parkesh, R.; Pushechnikov, A., Structural Enrichment of Hts Compounds from Available Commercial Libraries. *MedChemComm* **2012**, *3*, 571-579.
9. Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K. C., Assessment of Chemical Libraries for Their Druggability. *Comput. Biol. Chem.* **2005**, *29*, 55-67.
10. Verheij, H. J., Leadlikeness and Structural Diversity of Synthetic Screening Libraries. *Mol. Divers.* **2006**, *10*, 377-88.
11. Shang, J.; Sun, H.; Liu, H.; Chen, F.; Tian, S.; Pan, P.; Li, D.; Kong, D.; Hou, T., Comparative Analyses of Structural Features and Scaffold Diversity for Purchasable Compound Libraries. *J. Cheminformatics* **2017**, *9*, 25.
12. Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W.; Tomkinson, N. P., Designing Novel Building Blocks Is an Overlooked Strategy to Improve Compound Quality. *Drug Discov. Today* **2015**, *20*, 11-7.
13. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S., Probing the Bioactivity-Relevant Chemical Space of Robust Reactions and Common Molecular Building Blocks. *J. Chem. Inf. Model.* **2012**, *52*, 1167-78.
14. Gvk Biosciences Private Limited. Plot No. 28 a, Ida Nacharam, Hyderabad 500076, India.
15. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J., Drugbank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668-72.
16. Higuieruelo, A. P.; Schreyer, A.; Bickerton, G. R.; Pitt, W. R.; Groom, C. R.; Blundell, T. L., Atomic Interactions and Profile of Small Molecules Disrupting Protein-Protein Interfaces: The Timbal Database. *Chem. Biol. Drug. Des.* **2009**, *74*, 457-67.
17. Kalliokoski, T., Price-Focused Analysis of Commercially Available Building Blocks for Combinatorial Library Synthesis. *ACS Comb. Sci.* **2015**, *17*, 600-7.
18. Emolecules, Inc. <https://www.emolecules.com/>. (14.05.2020),
19. Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Gavrylenko, K.; Horvath, D.; Klimchuk, O.; Oksiuta, O.; Marcou, G.; Varnek, A., Synthi: A New Open-Source Tool for Synthon-Based Library Design. *J. Chem. Inf. Model.* **2021**, DOI: 10.1021/acs.jcim.1c00754.
20. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R., ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930-D940.
21. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D., Isida Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010**, *29*, 855-68.
22. Bishop, C. M.; Svensén, M.; Williams, C. K. I., Gtm: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10*, 215-234.
23. Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J. L.; Varnek, A., Mapping of the Available Chemical Space Versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018**, *13*, 540-554.
24. Horvath, D.; Marcou, G.; Varnek, A., Generative Topographic Mapping in Drug Design. *Drug Discov. Today Technol.* **2019**, 32-33, 99-107.
25. Lin, A.; Beck, B.; Horvath, D.; Marcou, G.; Varnek, A., Diversifying Chemical Libraries with Generative Topographic Mapping. *J. Comput. Aided Mol. Des.* **2020**, *34*, 805-815.



26. Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A., Np Navigator: A New Look at the Natural Product Chemical Space. *Mol. Inform.*, doi:10.1002/minf.202100068 **2021**, 40, 2100068.
27. Gaspar, H. A.; Baskin, II; Marcou, G.; Horvath, D.; Varnek, A., Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J.Chem.Inf.Model.* **2015**, 55, 84-94.
28. ChemAxon. Standardizer, C Version 5.12, ChemAxon, Ltd: Budapest, Hungary **2012**.
29. Tanimoto, T., An Elementary Mathematical Theory of Classification and Prediction. *Internal IBM Technical Report.* **1958**, 8.
30. Kireeva, N.; Baskin, II; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A., Generative Topographic Mapping (Gtm): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **2012**, 31, 301-12.
31. Gaspar, H. A.; Baskin, II; Marcou, G.; Horvath, D.; Varnek, A., Gtm-Based Qsar Models and Their Applicability Domains. *Mol. Inform.* **2015**, 34, 348-56.
32. Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A., Multi-Task Generative Topographic Mapping in Virtual Screening. *J. Comput. Aided Mol. Des.* **2019**, 33, 331-343.
33. Vantourout, J. C.; Law, R. P.; Isidro-Llobet, A.; Atkinson, S. J.; Watson, A. J. B., Chan–Evans–Lam Amination of Boronic Acid Pinacol (Bpin) Esters: Overcoming the Aryl Amine Problem. *The Journal of Organic Chemistry* **2016**, 81, 3942-3950.
34. Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A., Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J.Chem.Inf.Model.* **2019**, 59, 564-572.
35. Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D., Mappability of Drug-Like Space: Towards a Polypharmacologically Competent Map of Drug-Relevant Compounds. *J. Comput. Aided Mol. Des.* **2015**, 29, 1087-108.
36. Brown, D. G.; Bostrom, J., Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, 59, 4443-58.
37. Cooper, T. W.; Campbell, I. B.; Macdonald, S. J., Factors Determining the Selection of Organic Reactions by Medicinal Chemists and the Use of These Reactions in Arrays (Small Focused Libraries). *Angew. Chem., Int. Ed.* **2010**, 49, 8082-91.
38. Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S., Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, 23, 101681.
39. Lovering, F.; Bikker, J.; Humblet, C., Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, 52, 6752-6.
40. Lovering, F., Escape from Flatland 2: Complexity and Promiscuity. *MedChemComm* **2013**, 4, 515-519.
41. Tomberg, A.; Bostrom, J., Can Easy Chemistry Produce Complex, Diverse, and Novel Molecules? *Drug Discov. Today* **2020**, 25, 2174-2181.
42. Ward, R. A.; Kettle, J. G., Systematic Enumeration of Heteroaromatic Ring Systems as Reagents for Use in Medicinal Chemistry. *J. Med. Chem.* **2011**, 54, 4670-7.
43. Buskes, M. J.; Blanco, M. J., Impact of Cross-Coupling Reactions in Drug Discovery and Development. *Molecules* **2020**, 25 (15), 3493.
44. Nicolaou, C. A.; Watson, I. A.; LeMasters, M.; Masquelin, T.; Wang, J., Context Aware Data-Driven Retrosynthetic Analysis. *J.Chem.Inf.Model.* **2020**, 60, 2728-2738.

45. Vasudevan, A.; Bogdan, A. R.; Koolman, H. F.; Wang, Y.; Djuric, S. W. Chapter One - Enabling Chemistry Technologies and Parallel Synthesis—Accelerators of Drug Discovery Programmes. In *Progress in Medicinal Chemistry*, Witty, D. R.; Cox, B., Eds.; Elsevier: 2017; Vol. 56, 1-35.
46. Li, J.; Ballmer, S. G.; Gillis, E. P.; Fujii, S.; Schmidt, M. J.; Palazzolo, A. M.; Lehmann, J. W.; Morehouse, G. F.; Burke, M. D., Synthesis of Many Different Types of Organic Small Molecules Using One Automated Process. *Science* **2015**, *347*, 1221-6.
47. Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L. C.; Schneeweis, J.; Berritt, S.; Shi, Z. C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D., Organic Chemistry. Nanomole-Scale High-Throughput Chemistry for the Synthesis of Complex Molecules. *Science* **2015**, *347*, 49-53.
48. Perera, D.; Tucker, J. W.; Brahmabhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W., A Platform for Automated Nanomole-Scale Reaction Screening and Micromole-Scale Synthesis in Flow. *Science* **2018**, *359*, 429-434.
49. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G., Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science* **2018**, *360*, 186-190.
50. Roughley, S. D.; Jordan, A. M., The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54*, 3451-79.
51. Afanasyev, O. I.; Kuchuk, E.; Usanov, D. L.; Chusov, D., Reductive Amination in the Synthesis of Pharmaceuticals. *Chem. Rev.* **2019**, *119*, 11857-11911.
52. Epifanov, M.; Mo, J. Y.; Dubois, R.; Yu, H.; Sammis, G. M., One-Pot Deoxygenation and Substitution of Alcohols Mediated by Sulfuryl Fluoride. *J. Org. Chem.* **2021**, *86*, 3768-3777.
53. Figlus, M.; Wellaway, N.; Cooper, A. W.; Sollis, S. L.; Hartley, R. C., Synthesis of Arrays Using Low Molecular Weight Mpeg-Assisted Mitsunobu Reaction. *ACS Comb. Sci.* **2011**, *13*, 280-5.
54. Huang, H.; Kang, J. Y., Mitsunobu Reaction Using Basic Amines as Pronucleophiles. *J. Org. Chem.* **2017**, *82*, 6604-6614.
55. Hamid, M. H.; Allen, C. L.; Lamb, G. W.; Maxwell, A. C.; Maytum, H. C.; Watson, A. J.; Williams, J. M., Ruthenium-Catalyzed N-Alkylation of Amines and Sulfonamides Using Borrowing Hydrogen Methodology. *J. Am. Chem. Soc.* **2009**, *131*, 1766-74.
56. Wright, S. W.; Hallstrom, K. N., A Convenient Preparation of Heteroaryl Sulfonamides and Sulfonyl Fluorides from Heteroaryl Thiols. *J. Org. Chem.* **2006**, *71*, 1080-4.
57. Dong, J.; Krasnova, L.; Finn, M. G.; Sharpless, K. B., Sulfur(VI) Fluoride Exchange (Sufex): Another Good Reaction for Click Chemistry. *Angew. Chem., Int. Ed.* **2014**, *53*, 9430-48.
58. Grygorenko, O. O.; Biitseva, A. V.; Zherish, S., Amino Sulfonic Acids, Peptidosulfonamides and Other Related Compounds. *Tetrahedron* **2018**, *74*, 1355-1421.
59. Mykhalchuk, V. L.; Yarmolchuk, V. S.; Doroschuk, R. O.; Tolmachev, A. A.; Grygorenko, O. O., [3+2] Cycloaddition of an Azomethyne Ylide and Vinyl Sulfonyl Fluorides — an Approach to Pyrrolidine-3-Sulfonyl Fluorides. *Eur. J. Org. Chem.* **2018**, *2018*, 2870-2876.
60. Sokolov, A.; Golovach, S.; Kozlinsky, I.; Dolia, K.; Tolmachev, A. A.; Kuchkovska, Y.; Grygorenko, O. O., Diastereoselective Synthesis of Cyclic Sp<sup>3</sup>-Enriched Cis-B-Alkoxysulfonyl Chlorides. *Synthesis* **2019**, *51*, 848-858.
61. Tolmachova, K. A.; Moroz, Y. S.; Konovets, A.; Platonov, M. O.; Vasylenko, O. V.; Borysko, P.; Zozulya, S.; Gryniukova, A.; Bogolubsky, A. V.; Pipko, S.; Mykhailiuk, P. K.; Brovarets, V. S.; Grygorenko, O. O., (Chlorosulfonyl)Benzenesulfonyl Fluorides-Versatile

Building Blocks for Combinatorial Chemistry: Design, Synthesis and Evaluation of a Covalent Inhibitor Library. *ACS Comb. Sci.* **2018**, *20*, 672-680.

62. Kokhan, S. O.; Valter, Y. B.; Tymtsunik, A. V.; Komarov, I. V.; Grygorenko, O. O., 3-Carboxy-/3-Aminobicyclo[1.1.1]Pentane-Derived Sulfonamides and Sulfonyl Fluorides - Advanced Bifunctional Reagents for Organic Synthesis and Drug Discovery. *Eur. J. Org. Chem.* **2020**, *2020*, 2210-2216.

63. Helal, C. J.; Bundesmann, M.; Hammond, S.; Holmstrom, M.; Klug-McLeod, J.; Lefker, B. A.; McLeod, D.; Subramanyam, C.; Zakaryants, O.; Sakata, S., Quick Building Blocks (Qbb): An Innovative and Efficient Business Model to Speed Medicinal Chemistry Analog Synthesis. *ACS Med. Chem. Lett.* **2019**, *10*, 1104-1109.

64. Collins, K. D.; Glorius, F., A Robustness Screen for the Rapid Assessment of Chemical Reactions. *Nat. Chem.* **2013**, *5*, 597-601.

65. Fujiwara, Y.; Dixon, J. A.; O'Hara, F.; Funder, E. D.; Dixon, D. D.; Rodriguez, R. A.; Baxter, R. D.; Herle, B.; Sach, N.; Collins, M. R.; Ishihara, Y.; Baran, P. S., Practical and Innate Carbon-Hydrogen Functionalization of Heterocycles. *Nature* **2012**, *492*, 95-9.

66. Crisenza, G. E. M.; Melchiorre, P., Chemistry Glows Green with Photoredox Catalysis. *Nat. Commun.* **2020**, *11*, 803.

67. Meng, G.; Guo, T.; Ma, T.; Zhang, J.; Shen, Y.; Sharpless, K. B.; Dong, J., Modular Click Chemistry Libraries for Functional Screens Using a Diazotizing Reagent. *Nature* **2019**, *574*, 86-89.

68. Barrow, A. S.; Smedley, C. J.; Zheng, Q.; Li, S.; Dong, J.; Moses, J. E., The Growing Applications of Sufex Click Chemistry. *Chem. Soc. Rev.* **2019**, *48*, 4731-4758.

69. Cernak, T.; Dykstra, K. D.; Tyagarajan, S.; Vachal, P.; Krska, S. W., The Medicinal Chemist's Toolbox for Late Stage Functionalization of Drug-Like Molecules. *Chem. Soc. Rev.* **2016**, *45*, 546-76.

70. Proctor, R. S. J.; Phipps, R. J., Recent Advances in Minisci-Type Reactions. *Angew. Chem., Int. Ed.* **2019**, *58*, 13666-13699.

71. Zhang, G.; Howe, M.; Aldrich, C. C., Spirocyclic and Bicyclic 8-Nitrobenzothiazinones for Tuberculosis with Improved Physicochemical and Pharmacokinetic Properties. *ACS Med. Chem. Lett.* **2019**, *10*, 348-351.

72. Manka, J. T.; Rodriguez, A. L.; Morrison, R. D.; Venable, D. F.; Cho, H. P.; Blobaum, A. L.; Daniels, J. S.; Niswender, C. M.; Conn, P. J.; Lindsley, C. W.; Emmitte, K. A., Octahydropyrrolo[3,4-C]Pyrrole Negative Allosteric Modulators of Mglu1. *Bioorg. Med. Chem. Lett.* **2013**, *23*, 5091-6.

73. Grygorenko, O. O.; Radchenko, D. S.; Volochnyuk, D. M.; Tolmachev, A. A.; Komarov, I. V., Bicyclic Conformationally Restricted Diamines. *Chem. Rev.* **2011**, *111*, 5506-68.

74. Cahard, D.; Ma, J. A., *Emerging Fluorinated Motifs: Synthesis, Properties, and Applications*. John Wiley & Sons, Ltd: Chichester, UK: 2020.

75. Meanwell, N. A., Fluorine and Fluorinated Motifs in the Design and Application of Bioisosteres for Drug Design. *J. Med. Chem.* **2018**, *61*, 5822-5880.

76. Mader, P.; Kattner, L., Sulfoximines as Rising Stars in Modern Drug Discovery? Current Status and Perspective on an Emerging Functional Group in Medicinal Chemistry. *J. Med. Chem.* **2020**, *63*, 14243-14275.

77. Finkbeiner, P.; Hehn, J. P.; Gnam, C., Phosphine Oxides from a Medicinal Chemist's Perspective: Physicochemical and in Vitro Parameters Relevant for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 7081-7107.

78. Ramesh, R.; Reddy, D. S., Quest for Novel Chemical Entities through Incorporation of Silicon in Drug Scaffolds. *J. Med. Chem.* **2018**, *61*, 3779-3798.
79. Mykhailiuk, P. K., Saturated Bioisosteres of Benzene: Where to Go Next? *Org. Biomol. Chem.* **2019**, *17*, 2839-2849.

SYNOPSIS (Word Style "SN\_Synopsis\_TOC"). If you are submitting your paper to a journal that requires a synopsis, see the journal's Instructions for Authors for details.

